Behaviorist Thinking in Judgments of Wrongness, Punishment, and Blame

Julian de Freitas[*1] & Samuel G. B. Johnson[*2]
(julian.defreitas@psy.ox.ac.uk, samuel.johnson@yale.edu)
[1]Department of Experimental Psychology, Oxford University; [2]Department of Psychology, Yale University
***Authors contributed equally to this work**

Abstract

Moral judgment depends upon inferences about agents' beliefs, desires, and intentions. Here, we argue that in addition to these factors, people take into account the moral *optimality* of an action. Three experiments show that even agents who are ignorant about the nature of their moral decisions are held accountable for the quality of their decision—a kind of *behaviorist thinking*, in that such reasoning bypasses the agent's mental states. In particular, whereas optimal choices are seen as more praiseworthy than suboptimal choices, decision quality has no further effect on moral judgments—a highly suboptimal choice is seen as no worse than a marginally suboptimal choice. These effects held up for judgments of wrongness and punishment (Experiment 1), positive and negative outcomes (Experiment 2), and agents with positive and negative intentions (Experiment 3). We argue that these results reflect a broader tendency to irresistibly apply the Efficiency Principle when explaining behavior.

Keywords: Moral judgment; theory of mind; causal reasoning; intentionality; lay decision theory.

Introduction

We judge others' actions on the basis of what they were thinking. If Megan cheats on an exam, we wish to know why she did it. If Joey made a hurtful remark, our opinion of him depends on whether he knew that his remark would be taken that way. We hold others accountable for their actions largely as a function of their intent—the joint product of their knowledge and goals (Dennett, 1987). This link between mindreading and moral judgment has been extensively studied in psychology (e.g., Cushman, 2008) and is enshrined in the law (Mikhail, 2007).

Yet, in some contexts, some of the functions of mindreading can be accomplished using mechanisms that do not actually involve reasoning about beliefs and desires. We might infer that the car to our right will change lanes on the basis of the assumed beliefs and desires of the car's driver; but more likely, we simply look at the car's blinker and read the car's future behavior off of the world directly. More generally, we often can predict behavior accurately by assuming that agents will behave *optimally* relative to assumed common goals (such as not colliding with other cars) and assumed common knowledge (the geometry of driving). This is known as taking the *teleological stance* (Dennett, 1987).

Infants can use situational information to predict the behavior of animate agents, well before they can perform explicit mental state computations. For example, infants are surprised if agents take an inefficient path between two locations; if an agent does take an inefficient path around an occluder, they expect an obstacle to be revealed behind the occluder (Gergely & Csibra, 2003). That is, infants apply the *Efficiency Principle*, assuming that agents behave optimally relative to situational constraints.

The Efficiency Principle is a fundamental principle which even adults use for predicting and explaining behavior (Baker, Saxe, & Tenenbaum, 2009; Johnson & Rips, 2015). Indeed, this principle is so ingrained in cognition that it can lead to *behaviorist thinking*, wherein even agents who are ignorant of their situational constraints are assumed to behave optimally relative to them. In one study (Johnson & Rips, 2014), participants read about agents making decisions under uncertainty. For example, Jill is deciding which of three shampoos to buy, wanting to make her hair smell like apples. She believes that all three brands have the same likelihood of achieving this goal, but in fact the three brands differ in quality—one has a 70% chance, one has a 50% chance, and one has a 30% chance. In reality, of course, Jill is equally likely to choose any of the three options, since, due to her ignorance, she has no basis for choosing one over another.

But participants seem to ignore Jill's ignorance, and use knowledge about the world to predict her decision: They predict that she will choose the optimal (70%) option rather than the suboptimal (50% or 30%) options. Most strikingly, even though participants judge the optimal (70%) option to be her most likely choice, they believe that the suboptimal 50% and 30% options are equally likely—that is, they use the Efficiency Principle. Even though the probability of the outcome differs just as much between the 70% and 50% options as between the 50% and 30% options, only the 70% option is optimal or efficient. Like infants, adults appear to use efficiency-based behaviorist thinking in understanding behavior.

In the current paper, we tested whether people use behaviorist thinking in moral judgment. If moral judgment depends on *mindreading* as such, they may not, since behaviorist thinking is an alternative, non-mentalistic strategy for understanding behavior. But if moral judgment instead depends on whatever tools we use for predicting and explaining behavior, behaviorist principles might seep into the moral realm.

This question was examined in three experiments. In all cases, agents made morally laden decisions in which (a) three potential options differed in quality, but (b) agents falsely believed that the options were equivalent. If people use the same behaviorist logic in moral judgment that they use to understand behavior in other cases, we would expect them to base their moral judgments on the quality of the agent's choice, even though the agents were

ignorant. Further, if this behaviorist thinking is a consequence of the Efficiency Principle, we would expect participants to give more lenient moral judgments only if an agent made an optimal choice; we would not expect moral judgments to differ between suboptimal choices that varied equally in quality.

Experiment 1

In Experiment 1, we tested for moral behaviorism in judgments of wrongness and of punishment. To vary the actual quality of the choice while leaving the agent's beliefs and intentions constant, participants were told that the agent had a false belief. One vignette read (material in brackets varied across vignettes, between-subjects):

A doctor working in a hospital has a patient who is having hearing problems. This patient has three, and only three, treatment options. The doctor believes that all treatment options have a 70% chance of giving the patient a full, successful recovery. But in fact the doctor's belief is wrong. Actually:

1) If she gives the patient treatment LPN, there is a 70% chance the patient will have a full recovery.

2) If she gives the patient treatment PTY, there is a 50% chance the patient will have a full recovery.

3) If she gives the patient treatment NRW, there is a 30% chance the patient will have a full recovery.

The doctor chooses treatment [LPN/PTY/NRW], and the patient does not recover at all. The patient now has permanent hearing loss.

That is, the moral agent (here, the doctor) had three choices, which she believed to be of equal quality, but which in fact varied. We refer to the 70% option as *Best*, the 50% option as *Middle*, and the 30% option as *Worst*. Note that the probabilistic difference between Best and Middle is the same as that between Middle and Worst, but that only Best maximizes the probability of the outcome. That is, Best is the optimal decision, even though the doctor has no way of knowing this.

Participants were asked, given the negative outcome, to rate either the wrongness of the agent's action or the extent to which the agent should be punished. One possibility is that only punishment judgments would be influenced by the probability of the outcome, because prior work has found that the degree to which an agent is deemed punishable is primarily a function of the agent's causal contribution to that outcome (Cushman, 2008). To the extent that the agent made the outcome more probable, she should be judged more causal (Cheng, 1997; Johnson & Rips, 2013), and hence more deserving of punishment. In that case, agents should be seen as more deserving of punishment when they choose Worst than when they choose Middle, and more when they choose Middle than when they choose Best. However, because judgments of wrongness tend to track intentions rather than causation (Cushman, 2008), we would expect no differences in wrongness judgments between agents who choose Best, Middle, or Worst.

Another possibility, however, is that people use behaviorist thinking when judging both punishment and wrongness, because people apply the Efficiency Principle to all agents, even agents who lack critical information about their decision situation (Johnson & Rips, 2014). That is, if people have a tacit expectation that moral agents behave optimally, then judgments of both wrongness and punishment could track the optimality of an action. Thus, we would always expect more favorable moral judgments when agents choose Best rather than Middle, since Best is optimal while Middle is suboptimal, but we would expect little or no difference between choosing Middle and Worst, since both are suboptimal.

Method

We recruited 336 participants from Amazon Mechanical Turk; 80 were excluded because they incorrectly answered one or more check questions.

Participants were assigned to one of eight vignettes (doctor, farmer, contractor, programmer, pilot, paramedic, CEO, or investment broker), formatted as above. The agent was always ignorant about the quality of the Best (70% chance of a positive outcome), Middle (50% chance), and Worst (30% chance) options. After choosing either the Best, Middle, or Worst option (between-subjects), the decision always led to a negative outcome.

On the same page, participants answered either a question about wrongness (e.g., "How wrong was the doctor's behavior") or punishment ("How much should the doctor be punished"), on a continuous scale (1 = "not at all", 4 = "somewhat", 7 = "very much"). Thus, the design of the experiment was 8 (vignette) x 3 (choice: Best, Middle, or Worst) x 2 (DV: wrongness or punishment), with all factors between-subjects.

On the next page, participants were asked two check questions, to ensure they had understood the vignette. One question concerned the agent's knowledge (e.g., "Did the doctor know about the actual chances of success for each of the options?") and the other question concerned the agent's choice ("What was the actual chance of success for the option which the doctor chose?"). Because our hypotheses are predicated on the assumption that participants understand the agent's knowledge state and choice, any participant who made an error on either question was excluded from analysis.

Results and Discussion

In all vignettes, agents made their decision under the false belief that all decision options were of equal quality. Thus, the agents' mental states were identical across the three conditions of the experiment. Further, a negative result occurred in all cases, so the outcome was also identical across conditions. Yet, Figure 1 shows that participants' judgments of both wrongness and punishment differed depending on the agent's choice.

For wrongness judgments, the agents who chose the (optimal) Best option were judged more leniently than

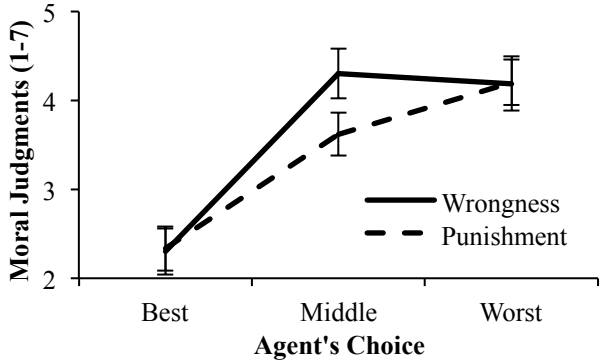

Figure 1: Results of Experiment 1. Bars represent 1 SE.

those who chose the (suboptimal) Middle option [M = 2.30, SD = 1.71 vs. M = 4.30, SD = 1.90; $t(87) = -5.22$, $p < .001$, $d = -1.11$, $BF_{10} > 1000$]. Likewise, agents who chose the Best option were judged less deserving of punishment than those who chose the Middle option [M = 2.33, SD = 1.54 vs. M = 3.62, SD = 1.48; $t(74) = -3.71$, $p < .001$, $d = -0.85$, $BF_{10} = 64.8$].[1] These large differences held up despite the agents' ignorance and despite the ultimately negative outcome, indicating that people base their moral judgments not only on the agents' mental states and the extremity of the outcome, but also on the agent's optimality. Thus, people appear to use behaviorist thinking in their moral evaluations.

However, participants' moral judgments did not differ between different suboptimal choices. Wrongness judgments were no less harsh when agents chose the Middle option than when they chose the Worst option [M = 4.30, SD = 1.90 vs. M = 4.19, SD = 1.97; $t(86) = 0.28$, $p = .78$, $d = 0.06$, $BF_{01} = 5.90$], nor were punishment judgments [M = 3.62, SD = 1.48 vs. M = 4.20, SD = 1.79; $t(84) = -1.61$, $p = .11$, $d = -0.35$, $BF_{01} = 1.83$]. These null effects—with evidence favoring the null hypothesis, according to the Bayes Factor analyses—are remarkable in light of the very large differences between the Best and Middle conditions, which had just as large of a difference in efficacy (a 70% vs. 50% chance of a positive outcome) as the difference between the Middle and Worst conditions (50% vs. 30%). Most importantly, these null effects reveal the same pattern of results as Johnson and Rips (2014) found in people's predictions and explanations for (non-moral) decisions. Finding a difference between optimal and suboptimal choices, but no difference between two different suboptimal choices of different efficacy suggests that participants' behaviorist thinking is rooted in efficiency-based reasoning, for which optimality is the key constraint.

A skeptic might raise two concerns about this interpretation. First, rather than blaming agents directly for choosing suboptimally, participants could instead have blamed the agents for their ignorance itself. This would not necessarily constitute behaviorist thinking, since agents' ignorance could be thought of as a failure to act with due diligence. Experiment 2 addressed this concern.

A second way that these findings could reflect mental state inferences would be if participants were using the agents' actions to infer their goals. If an agent chooses suboptimally, perhaps she did not want a positive outcome to occur. Experiment 3 addressed this concern.

Experiment 2

We had two goals in Experiment 2. First, we addressed a possible concern about Experiment 1—that participants may have held the agents accountable not for their suboptimal choice, but for the ignorance that led to that choice. Although this possibility would not *a priori* predict the pattern of results we found in Experiment 1 (since the agents were ignorant in all conditions), perhaps participants were willing to forgive the agents for their ignorance when they were fortunate enough to have chosen optimally (a possibility reminiscent of the outcome bias; Baron & Hershey, 1988). To avoid this concern, we added a sentence to each vignette indicating that the agent had arrived at their (mistaken) opinion about the decision options after undertaking extensive research. Thus, the agents in Experiment 2 have done their due diligence to understand their decisions.

Second, we wanted to test whether people apply behaviorist thinking in light of both positive and negative outcomes. Because people tend to make harsher moral evaluations in light of negative outcomes than positive outcomes (Baron & Hershey, 1988), perhaps all's well that ends well—the quality of the agent's decision may be seen as irrelevant when the outcome is positive.

Method

We recruited 267 participants from Amazon Mechanical Turk; 67 were excluded because they incorrectly answered one or more check questions.

Vignettes were the same as those in Experiment 1, except that they indicated that the agent had undertaken extensive research before making her decision. For example, in the doctor vignette, the third sentence from Experiment 1 was replaced with:

> *Based on many articles that the doctor has carefully read in respected medical journals, she truly believes that all three options have a 70% chance of giving the patient a full, successful recovery.*

The only other difference in the vignettes was that the outcome could be negative, as in Experiment 1 (e.g., "…the patient does not recover at all. The patient now has permanent hearing loss") or positive ("…the patient

[1] All t-tests reported in this paper are supplemented with a Bayes Factor (BF) with a scale factor of 1 (Rouder, Speckman, Sun, Morey, & Iverson, 2009). For example, "$BF_{10} = 4.0$" means that the data would be 4 times likelier under the alternative hypothesis than under the null hypothesis, giving reason to reject the null hypothesis. However, BFs can also quantify evidence in favor of a null hypothesis; "$BF_{01} = 6.0$" means that the data would be 6 times likelier under the null than under the alternative, giving reason to accept the null hypothesis.

recovers. The patient has no permanent hearing loss"). The design was thus 8 (vignette) x 3 (choice: Best, Middle, or Worst) x 2 (outcome: positive or negative), with all factors between-subjects.

On the same page, participants answered, "What does the doctor deserve to receive for her behavior?" on a scale from 1 ("extreme blame") to 9 ("extreme praise"). On the following page, participants answered the same check questions used in Experiment 1.

Results and Discussion

Figure 2 shows that for both positive and negative outcomes, participants again gave more positive moral judgments for agents choosing the (optimal) Best option compared to the (suboptimal) Middle or Worst options. Specifically, participants assigned more moral praise to agents who chose Best than those who chose Middle, for positive outcomes [M = 6.71, SD = 1.42 vs. M = 5.22, SD = 1.68; $t(61) = 3.80$, $p < .001$, $d = 0.96$, $BF_{10} = 78.0$] and for negative outcomes [M = 4.37, SD = 1.54 vs. M = 3.25, SD = 1.40; $t(64) = 3.08$, $p < .001$, $d = 0.76$, $BF_{10} = 11.3$]. That is, participants attended not only to the outcome and to the agent's mental states, but also to the quality of the agent's choice. This was true even though the agents were ignorant of the true quality of the possible options, and had taken extensive efforts to be knowledgeable.

Yet, just as in Experiment 1, the Middle and Worst options did not differ, even though the probability of the outcome differed just as much between Middle and Worst as between Best and Middle. Agents who chose Middle were rated no more praiseworthy than those who chose Worse, whether the outcome was positive [M = 5.22, SD = 1.68 vs. M = 4.65, SD = 1.76; $t(64) = 1.35$, $p = .18$, d = 0.33, $BF_{01} = 2.4$] or negative [M = 3.25, SD = 1.40 vs. M = 2.97, SD = 1.48; $t(71) = 0.82$, $p = .42$, $d = 0.19$, BF_{01} = 4.1]. Thus, once again, participants appear to be basing their moral evaluations off of optimality, rather than the probability of the outcome, since Best differed in optimality from Middle and Worst, whereas all three options differed equally in probability. This suggests that participants are applying the Efficiency Principle to the ignorant moral decision-makers.

Participants also used the outcome (positive or negative) in their moral judgments, assigning more praise for positive outcomes than for negative outcomes [M = 5.49, SD = 1.83 vs. M = 3.48, SD = 1.57; $t(198) = 8.38$, $p < .001$, $d = 1.19$, $BF_{10} > 1000$], even given precisely the same choice on the part of the agent, as shown in Figure 2. This is consistent with Baron and Hershey's (1988) demonstrations of the outcome bias. Given the robustness of the outcome bias in the moral psychology literature, it is worth noting that the effect of optimality (d = 0.96 for positive outcomes and d = 0.76 for negative outcomes) approached the effect size of the outcome bias (d = 1.19; see also Figure 2 for means).

These results cast doubt on the possibility that people are holding the agents accountable because of their ignorance, rather than because of their choice. Because the agents in Experiment 2 took appropriate steps to secure their knowledge, it is unlikely that participants believed their ignorance to be a moral failing.

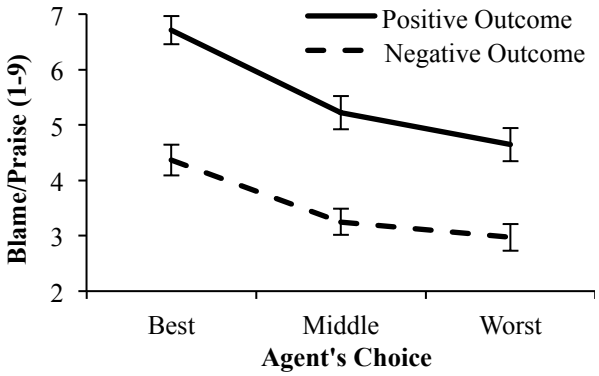

Figure 2: Results of Experiment 2. Bars represent 1 SE.

Experiment 3

In Experiment 3, we addressed the possibility that participants in Experiments 1 and 2 were interpreting the agents' choices as reflecting their goals—optimal decisions could have signaled a positive or beneficent intention, whereas suboptimal decisions could have signaled a negative or malevolent intention. Once again, it is unclear on this account why a highly suboptimal choice would not signal a worse intention than a moderately suboptimal choice. Nonetheless, we avoided this concern in Experiment 3 by specifying that the agents had either a positive intention (to achieve the morally mandated goal) or a negative intention (not to achieve it). If we continue to find similar effects across conditions, this would suggest that participants are relying on the efficiency of the agent's choice in a behaviorist manner, rather than relying on inferences about the agent's intentions.

Method

We recruited 335 participants from Amazon Mechanical Turk; 124 were excluded because they incorrectly answered one or more check questions.

The vignettes were the same as those used in Experiment 2, except that a sentence was added immediately before the third sentence, specifying either a positive intent (e.g., "The doctor intends to choose the best treatment option for her patient") or a negative intent ("The doctor does not intend to choose the best treatment option for her patient"). The design was thus 8 (vignette) x 3 (choice: Best, Middle, or Worst) x 2 (goal: positive or negative), with all factors between-subjects. The dependent measure was the same as in Experiment 2.

On the following page, participants answered the same check questions used in Experiment 2, in addition to a third question about the agent's intent (e.g., "Did the doctor intend to choose the best treatment option for her patient?"). As in the other experiments, participants were excluded for answering any check question incorrectly.

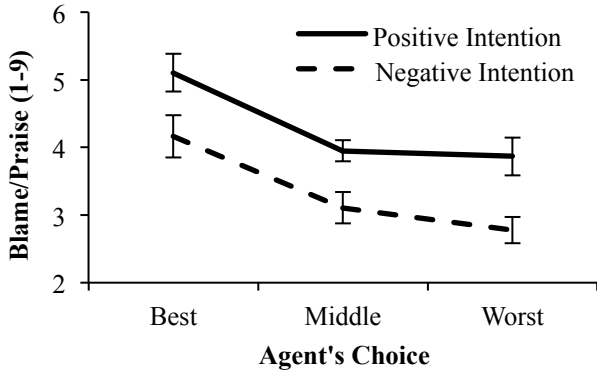

Figure 3: Results of Experiment 3. Bars represent 1 SE.

Results and Discussion

Whether the agents had positive or negative intentions, participants assigned more praise (or less blame) to the optimal than to the suboptimal moral decision-makers, replicating the same pattern as Experiments 1 and 2 (see Figure 3). The difference in blame/praise judgments between Best and Middle was robust both when the agent had positive intentions [M = 5.11, SD = 1.72 vs. M = 3.95, SD = 0.99; $t(76) = 3.66$, $p < .001$, $d = 0.83$, $BF_{10} = 56.4$] and when the agent had negative intentions [M = 4.17, SD = 1.72 vs. M = 3.11, SD = 3.11, SD = 1.41; $t(65) = 2.77$, $p = .007$, $d = 0.68$, $BF_{10} = 5.3$]. Yet, just as in previous experiments, there was no difference in praise judgments between agents who chose Middle and those who chose worst, either for positively intentioned [M = 3.95, SD = 0.99 vs. M = 3.87, SD = 1.55; $t(68) = 0.27$, $p = .78$, $d = 0.07$, $BF_{01} = 5.3$] or negatively intentioned agents [M = 3.11, SD = 1.41 vs. M = 2.78, SD = 1.17; $t(71) = 1.09$, $p = .28$, $d = 0.25$, $BF_{01} = 3.3$].

Just as Experiment 2 showed that the effect of optimal versus suboptimal choice is of nearly comparable size to the effect of positive versus negative outcomes, we can compare the effect of optimal choice to the effect of intention. Although there was a sizeable effect of positive versus negative intention [M = 4.33, SD = 1.53 vs. M = 3.30, SD = 1.53; $t(209) = 4.89$, $p < .001$, $d = 0.67$, $BF_{10} > 1000$], the effect size of Best versus Middle difference was, if anything, even larger (d = 0.83 for positive intentions and d = 0.68 for negative intentions).

Because similar effects of optimality were found in Experiment 3, even when the agents' positive or negative intentions were specified explicitly, it seems that the effect of optimality is not due to tacit inferences about the agents' goals. Together with Experiment 2, this seems to foreclose on the possibility that participants' judgments were driven by mental state inferences. Instead, participants appear to use the Efficiency Principle in moral judgment, just as they do when they explain behavior in other domains (Johnson & Rips, 2014).

General Discussion

Our moral judgments depend on how we explain the behavior of moral agents. We often understand others' behavior by *mindreading* or mental-state inference (e.g., Heider, 1958). But in other contexts, people seem to skip past inferences about mental states, instead predicting and explaining behavior based on information immediately perceptible in the world (Gergely & Csibra, 2003; Johnson & Rips, 2014). Do people also use such behaviorist or teleological thinking in moral judgment?

The current studies show that people do. People reliably gave more lenient moral judgments to agents who acted optimally (choosing the best of three options, with respect to a morally laden outcome) than to agents who acted suboptimally, even though the agents were ignorant about the quality of the options. Further, the *degree* of suboptimality did not matter (choosing the second-best or worst of the options). This pattern reflects use of the *Efficiency Principle*—the expectation that agents will act optimally relative to their situational constraints (Dennett, 1987). Yet, use of this principle was unjustified, because the agents were ignorant about these constraints. This pattern was replicated across six independent samples of participants (Experiments 1–3), and matches findings in previous studies of social cognition (Johnson & Rips, 2014). These findings both demonstrate behaviorist reasoning (since people base their moral judgments on the ignorant agents' choice) and pinpoint the Efficiency Principle as the mechanism (since people distinguish only between optimal and suboptimal choices).

Could a less radical, mindreading-based account also explain these findings? First, could participants have imbued the agent with knowledge of the options, despite the wording of the vignettes? It is unlikely that participants were misreading the materials, because participants were asked explicitly whether the agent was knowledgeable, and participants answering this question incorrectly were excluded from analysis. But could participants have been making a subtler mistake, confusing their own perspective with that of the agent (Birch & Bloom, 2007)? This possibility is also unlikely, because it would also predict a difference between different suboptimal choices (i.e., between *Middle* and *Worst*), which was not found in any of the experiments. In fact, Johnson and Rips (2014) compared people's inferences about knowledgeable and ignorant agents, and found that participants *do* distinguish between different suboptimal options when the agents are knowledgeable. (This finding also rules out the possibility that participants may have misinterpreted the probabilities because of a framing effect or probabilistic reasoning fallacy, since those fallacies would apply equally to reasoning about knowledgeable and ignorant agents.)

Our experiments also addressed two other ways that participants' moral judgments might be based on mental-state considerations. Experiment 2 speaks against the possibility that participants were blaming participants not

for their actions, but for their ignorance. In that study, the agents were said to have undertaken extensive research in coming to their beliefs, so that they had done their due diligence. Furthermore, we conducted an additional follow-up study, finding that participants applied the Efficiency Principle even when they were told that the probabilities were literally unknowable.

Experiment 3 addressed the possibility that people take the agents' choices to tacitly signal their intentions. Note, however, that even this explanation would still involve behaviorist thinking in that the ignorant agents could not have used the decision qualities in forming their intentions. Nonetheless, participants continued to make efficiency-based moral judgments even when the agents' intentions were explicitly stated in the vignettes.

A final concern is that our experiments may lack the statistical power to detect an effect of the Middle versus Worst choice, because this effect may be smaller than the effect of Best versus Middle. Although the Bayes Factors favoring the null hypothesis speak against this possibility, we further ruled out this concern with a follow-up study, using a within-subjects design and tripled sample size per cell. This high-power study revealed an identical pattern.

These results have implications for both moral psychology and for social cognition more broadly. In moral psychology, these results add efficiency to the set of heuristics that people use for moral reasoning (Sunstein, 2005). Further, they show that the role of behavior understanding in moral judgment goes beyond mindreading, to encompass the full range of tools that humans use for interpreting behavior (Dennett, 1987).

More generally, these results reinforce the fundamental place of the Efficiency Principle in social cognition. Efficiency considerations have been shown to play a role in infants' action understanding (Gergely & Csibra, 2003), in the perception of animacy (Gao & Scholl, 2011), in predicting agents' spatial trajectories (Baker et al., 2009), and in lay theories of decision-making (Johnson & Rips, 2013, 2014, 2015). The current results underscore the central explanatory role of efficiency, by extending this principle to moral judgment.

These results may also have widespread practical implications. Legal courts are often responsible for assessing culpability when a defendant was ignorant of some important aspect of their case. For instance, after the L'Aquila earthquake in Italy, six scientists were charged with manslaughter, even though it is impossible to accurately predict an earthquake (Cartlidge, 2009); and similar lawsuits have arisen in cases of negligence in gynecology and other medical practices, sometimes involving millions of dollars (e.g. Raghuveer, 2015).

The current work found that behaviorist thinking accounts for moral judgments in the complete absence of mental state information—an especially strong test of the behaviorism hypothesis. Future work should explore how and to what extent behaviorist thinking interacts with mindreading in moral judgments, and in what ways it extends to other domains of social cognition (see Johnson & Rips, 2014 for initial proposals).

Acknowledgments

We thank Fiery Cushman for detailed comments.

References

Baker, C.L., Saxe, R., & Tenenbaum, J.B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Baron, J., & Hershey, J.C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382–386.

Cartlidge, E. (2009). Updated: Appeals court overturns manslaughter convictions of six earthquake scientists. *Science News*.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Dennett, D.C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Gao, T., & Scholl, B.J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 669–684.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*, 287–292.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

Johnson, S.G.B., & Rips, L.J. (2013). Good decisions, good causes: Optimality as a constraint on attribution of causal responsibility. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Johnson, S.G.B., & Rips, L.J. (2014). Predicting behavior from the world: Naïve behaviorism in lay decision theory. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Johnson, S.G.B., & Rips, L.J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77,* 42–76.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*, 143–152.

Rahguveer, A. (2015) Jury awards $10.9M in malpractice case against Maumee OB-GYN. *NBC 24 News.*

Sunstein, C.R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*, 531–570.